Limits of Standard Deviation under constraints

Ian Hussey & Jamie Cummins

TODO

- Add references to prior work and relate this to them.
- Note that this does not account for the fact that means are typically rounded for reporting, whereas the algorithmic solution does.
- Note that this does not allow for application to mean scores from multi-item Likert scales, rather it assumes integer responses. The algorithmic solution does allow for this, although note that mean scores can be converted to sum scores by multiplying by k items - although too suffers from the impact of rounding.

Maximum SD of a truncated variable

Problem statement

Given n observations $x_1, x_2, \ldots, x_n$, each constrained to being integers in the interval $[\ell, u]$ and subject to a desired mean ($\bar{x}$) which itself lies in range $[\ell, u]$, we want to determine the distribution of these observations that yields the maximum sample Standard Deviation. These constraints correspond to those for data collected using Likert scales, whether single-item or the sum-scores of a multi-item scale, which are common in the social sciences.

Maximizing variance under constraints

For real numbers in $[\ell, u]$, the arrangement that maximizes variance (and therefore Standard Deviation) is to place as many observations as possible at the boundaries ℓ and u. Intuitively, a distribution with more "spread" around its mean exhibits higher variance.

Concretely, assume p observations are at u and q observations are at ℓ. If $p + q = n$ exactly satisfies the sum constraint $p\,u + q\,\ell = n\,\bar{x}$, then all values lie at the extremes. However, in many cases, one additional "middle" value m in $[\ell, u]$ is required to adjust the total sum to exactly $n\,\bar{x}$. In such scenarios:

- p observations are at u.
- $(n - p - 1)$ observations are at ℓ.
- Exactly one observation is at $m = \ell + \delta$, for some $\delta \in (0,\, u - \ell)$.

Determining the number of maximum values

To find how many observations should be at u, we solve

$$pu + (n - p)\,\ell \leq n\,\bar{x}, \tag{1}$$

which implies

$$p \leq \frac{n\,\bar{x} - n\,\ell}{u - \ell}. \tag{2}$$

Hence, the largest integer p we can use is

$$p = \left\lfloor \frac{n\,\bar{x} - n\,\ell}{u - \ell} \right\rfloor. \tag{3}$$

Calculating the leftover δ

After setting p values to u and $n - p$ values to ℓ, the current sum is

$$p\,u + (n - p)\,\ell = n\,\ell + p\,(u - \ell). \tag{4}$$

The difference between the required sum $n\,\bar{x}$ and the sum so far is

$$\delta = n\,\bar{x} - \big[p\,u + (n - p)\,\ell\big]. \tag{5}$$

Since δ must fit within $[\ell, u]$ for a single observation, it can be shown that that $\delta \in [\,0,\, u - \ell)$. If $\delta = 0$, then no middle value is needed. If $\delta > 0$, one observation is set to $\ell + \delta$.

Showing $\delta \in [0, u - \ell)$

By substituting the expression for the sum so far into the equation for δ:

$$\delta = n\bar{x} - \big[n\ell + p(u - \ell)\big]. \tag{6}$$

Expanding this gives:

$$\delta = n\bar{x} - n\ell - p(u - \ell). \tag{7}$$

Rearranging terms:

$$\delta = n(\bar{x} - \ell) - p(u - \ell). \tag{8}$$

Lower Bound for δ: The minimum value of δ occurs when p is maximized, that is, $p = \left\lfloor \frac{n(\bar{x}-\ell)}{u-\ell} \right\rfloor$. Substituting this into δ, we see that the leftover sum always satisfies $\delta \geq 0$, since the sum cannot exceed the required total.

Upper Bound for δ: The maximum value of δ occurs when one observation must adjust to make up the exact total. In this case, δ corresponds to the leftover amount of the total that does not perfectly divide into $pu + (n - p)\ell$, and since all values lie within $[\ell, u]$, the leftover is strictly less than $u - \ell$. Thus, $\delta \in [0, u - \ell)$.

If $\delta = 0$, no middle value is needed because the sum perfectly divides into pu and $(n - p)\ell$. Otherwise, one observation is assigned $\ell + \delta$.

Variance computation

The sum of squared deviations from the mean is:

- p observations at u:
$$p \times (u - \bar{x})^2. \tag{9}$$
- $(n - p - 1)$ observations at ℓ (if $\delta > 0$) or $(n - p)$ if $\delta = 0$:
$$(n - p - 1) \times (\ell - \bar{x})^2 \quad \text{or} \quad (n - p) \times (\ell - \bar{x})^2. \tag{10}$$
- One "middle" observation (only if $\delta > 0$):
$$(\ell + \delta - \bar{x})^2. \tag{11}$$

Summing these yields the total sum of squares SS:

$$\text{SS} = p \times (u - \bar{x})^2 + (n - p - 1) \times (\ell - \bar{x})^2 + (\ell + \delta - \bar{x})^2 \quad \text{if } \delta > 0, \tag{12}$$

or

$$\text{SS} = p \times (u - \bar{x})^2 + (n - p) \times (\ell - \bar{x})^2 \quad \text{if } \delta = 0. \tag{13}$$

The sample variance is

$$s^2 = \frac{\text{SS}}{n - 1}, \tag{14}$$

and the sample Standard Deviation is $s = \sqrt{s^2}$.

Because the distribution is "as extreme as possible," this arrangement yields a larger SS (and hence s^2) than any other arrangement that still satisfies $\sum x_i = n\,\bar{x}$. Thus, one obtains the maximum possible sample Standard Deviation (or variance) under the stated constraints.

Minimum SD of a truncated variable

Problem statement

Given n observations $x_1, x_2, \ldots, x_n$, each constrained to being integers in the interval $[\ell, u]$ and subject to a desired mean ($\bar{x}$) which itself lies in range $[\ell, u]$, we want to determine the distribution of these observations that yields the minimum sample Standard Deviation. Again, these constraints correspond to those for data collected using Likert scales, whether single-item or the sum-scores of a multi-item scale, which are common in the social sciences.

Minimising variance under constraints

Suppose we have a sample size n, a mean $\bar{x}$, and know that observations are constrained to be within the interval $[\ell, u]$, where ℓ and u are the minimum and maximum possible integer scores, respectively (e.g., Likert scale points).

We wish to minimize the sample Standard Deviation under these constraints. The key result is that, if it is feasible to have all observations equal to some integer $x \in [\ell, u]$ that matches the required mean, then

the Standard Deviation is 0, which is obviously minimal as variance (and therefore Standard Deviation) must be non-negative. Formally, if $n\bar{x}$ is an integer and $\ell \le \bar{x} \le u$ (with $\bar{x}$ itself an integer), one can set $x_1 = \cdots = x_n = \bar{x}$, yielding $s^2 = 0$ and therefore Standard Deviation also equal to 0. However, if $\bar{x}$ is not an integer or if $\bar{x}$ is outside the feasible range of a single integer value, the best strategy is to distribute observations across two adjacent integers. Concretely, let

$$L = \max\{\ell,\ \lfloor \bar{x} \rfloor\}, \tag{15}$$

be the largest integer not exceeding $\bar{x}$ (clamped to be at least ℓ), and let $r = n\bar{x} - nL$ be the "remainder" when forcing the mean to come from a baseline value L. In a purely theoretical (unbounded) setting, distributing r fraction of the observations at $L+1$ and the remainder at L yields the exact mean $\bar{x}$. Specifically, $n - r$ observations are placed at L and r observations are placed at $L+1$.

Since the scale is integer-valued, r must be an integer. If r falls between 0 and n and $\ell \le L \le u$ and $L + 1 \le u$, then the distribution

$$(L, L, \ldots, L, \underbrace{L+1, \ldots, L+1}_{r \text{ times}}), \tag{16}$$

matches the sum $n\bar{x}$ and satisfies the bounds.

No other arrangement of integer-valued observations with the same sum can yield a smaller variance, because any additional "spreading out" away from these two points (L and $L+1$) can only increase the sum of squared deviations from the mean.

Variance computation

If we take $n - r$ observations at L and r observations at $L + 1$, then the sample mean is forced to be $\bar{x} \approx L + r/n$.

The population variance of this two-point distribution can be expressed as:

$$s^2_{\text{pop}} = (1-p)(L - \bar{x})^2 + p((L+1) - \bar{x})^2, \tag{17}$$

where $p = \frac{r}{n}$ is the proportion of observations at $L + 1$.

We expand and simplify this to:

$$s^2_{\text{pop}} = \frac{r}{n}((L+1) - \bar{x})^2 + \left(1 - \frac{r}{n}\right)(L - \bar{x})^2. \tag{18}$$

Expanding further and substituting $\bar{x} = L + \frac{r}{n}$:

$$s^2_{\text{pop}} = \frac{r}{n}\left(\frac{n-r}{n}\right)^2 + \frac{n-r}{n}\left(-\frac{r}{n}\right)^2. \tag{19}$$

The sample variance (with denominator $n - 1$) is:

$$s^2 = \frac{n}{n-1} \cdot s^2_{\text{pop}}. \tag{20}$$

and the sample Standard Deviation is $s = \sqrt{s^2}$.

By substituting these values, we arrive at the closed-form solution for the minimal Standard Deviation. This approach confirms that the minimal-variance distribution uses at most two adjacent integer values, balanced according to the required mean.

Summary: The smallest-variance distribution for integer constraints and a given mean uses at most two adjacent integer values, with exact weights determined by the remainder $r = n\bar{x} - nL$.